Minimum Bayes-Risk Decoding for Statistical Machine Translation

Shankar Kumar and William Byrne *
Center for Language and Speech Processing, Johns Hopkins University,
3400 North Charles Street, Baltimore, MD, 21218, USA
{skumar,byrne}@jhu.edu

Abstract

We present Minimum Bayes-Risk (MBR) decoding for statistical machine translation. This statistical approach aims to minimize expected loss of translation errors under loss functions that measure translation performance. We describe a hierarchy of loss functions that incorporate different levels of linguistic information from word strings, word-to-word alignments from an MT system, and syntactic structure from parse-trees of source and target language sentences. We report the performance of the MBR decoders on a Chinese-to-English translation task. Our results show that MBR decoding can be used to tune statistical MT performance for specific loss functions.

1 Introduction

Statistical Machine Translation systems have achieved considerable progress in recent years as seen from their performance on international competitions in standard evaluation tasks (NIST, 2003). This rapid progress has been greatly facilitated by the development of automatic translation evaluation metrics such as BLEU score (Papineni et al., 2001), NIST score (Doddington, 2002) and Position Independent Word Error Rate (PER) (Och, 2002). However, given the many factors that influence translation quality, it is unlikely that we will find a single translation metric that will be able to judge all these factors. For example, the BLEU, NIST and the PER metrics, though effective, do not take into account explicit syntactic information when measuring translation quality.

Given that different Machine Translation (MT) evaluation metrics are useful for capturing different aspects of translation quality, it becomes desirable to create MT systems tuned with respect to each individual criterion. In contrast, the maximum likelihood techniques that underlie the decision processes of most current MT systems do not take into account these application specific goals. We apply the *Minimum Bayes-Risk* (MBR) techniques developed for automatic speech recognition (Goel and Byrne, 2000) and bitext word alignment for statistical MT (Kumar and Byrne, 2002), to the problem of building automatic MT systems tuned for specific metrics. This is a framework that can be used with statistical models of speech and language to develop decision processes optimized for specific loss functions.

We will show that MBR decoding can be applied to machine translation in two scenarios. Given an automatic MT metric, we design a loss function based on the metric and use MBR decoding to tune MT performance under the metric. We also show how MBR decoding can be used to incorporate syntactic structure into a statistical MT system by building specialized loss functions. These loss functions can use information from word strings, word-to-word alignments and parse-trees of the source sentence and its translation. In particular we describe the design of a *Bilingual Tree Loss Function* that can explicitly use syntactic structure for measuring translation quality. MBR decoding under this loss function allows us to integrate syntactic knowledge into a statistical MT system without building detailed models of linguistic features, and retraining the system from scratch.

We first present a hierarchy of loss functions for translation based on different levels of lexical and syntactic information from source and target language sentences. This hierarchy includes the loss functions useful in both situations where we intend to apply MBR decoding. We

*This work was supported by the National Science Foundation under Grant No. 0121285 and an ONR MURI Grant N00014-01-1-0685. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Minimum Bayes-Risk Decoding for Statistical Machine Translation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **John Hopkins University,Center for Language and Speech Processing,Department of Computer Science,Baltimore,MD,21218**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

then present the MBR framework for statistical machine translation under the various translation loss functions. We finally report the performance of MBR decoders optimized for each loss function.

2 Translation Loss Functions

We now introduce translation loss functions to measure the quality of automatically generated translations. Suppose we have a sentence F in a source language for which we have generated an automatic translation E' with word-to-word alignment A' relative to F. The word-to-word alignment A' specifies the words in the source sentence F that are aligned to each word in the translation E'. We wish to compare this automatic translation with a reference translation E with word-to-word alignment A relative to F.

We will now present a three-tier hierarchy of translation loss functions of the form $L((E', A'), (E, A); F)$ that measure (E', A') against (E, A). These loss functions will make use of different levels of information from word strings, MT alignments and syntactic structure from parse-trees of both the source and target strings as illustrated in the following table.

Loss Function	Functional Form
Lexical	$L(E, E')$
Target Language Parse-Tree	$L(T_E, T_{E'})$
Bilingual Parse-Tree	$L((T_E, A), (T_{E'}, A'); T_F)$

We start with an example of two competing English translations for a Chinese sentence (in Pinyin without tones), with their word-to-word alignments in Figure 1. The reference translation for the Chinese sentence with its word-to-word alignment is shown in Figure 2. In this section, we will show the computation of different loss functions for this example.

2.1 Lexical Loss Functions

The first class of loss functions uses no information about word alignments or parse-trees, so that $L((E', A'), (E, A); F)$ can be reduced to $L(E, E')$. We consider three loss functions in this category: The BLEU score (Papineni et al., 2001), word-error rate, and the position-independent word-error rate (Och, 2002). Another example of a loss function in this class is the MT-eval metric introduced in Melamed et al. (2003). A loss function of this type depends only on information from word strings.

BLEU score (Papineni et al., 2001) computes the geometric mean of the precision of n-grams of various lengths ($n \in \{1..N\}$) between a hypothesis and a reference translation, and includes a brevity penalty ($\gamma(E, E') \leq 1$) if the hypothesis is shorter than the reference. We use $N = 4$.

$$BLEU(E, E') = exp\left(\sum_{n=1}^{N} log\, \frac{p_n(E, E')}{N}\right) * \gamma(E, E'),$$

where $p_n(E, E')$ is the precision of n-grams in the hypothesis E'. The BLEU score is zero if any of the n-gram precisions $p_n(E, E')$ is zero for that sentence pair. We note that $0 \leq BLEU(E, E') \leq 1$. We derive a loss function from BLEU score as

$$L_{\text{BLEU}}(E, E') = 1 - BLEU(E, E').$$

Word Error Rate (WER) is the ratio of the string-edit distance between the reference and the hypothesis word strings to the number of words in the reference. String-edit distance is measured as the minimum number of edit operations needed to transform a word string to the other word string.

Position-independent Word Error Rate (PER) measures the minimum number of edit operations needed to transform a word string to any permutation of the other word string. The PER score (Och, 2002) is then computed as a ratio of this distance to the number of words in the reference word string.

2.2 Target Language Parse-Tree Loss Functions

The second class of translation loss functions uses information only from the parse-trees of the two translations, so that $L((E, A), (E', A'); F) = L(T_E, T_{E'})$. This loss function has no access to any information from the source sentence or the word alignments.

Examples of such loss functions are tree-edit distances between parse-trees, string-edit distances between event representation of parse-trees (Tang et al., 2002), and tree-kernels (Collins and Duffy, 2002). The computation of tree-edit distance involves an unconstrained alignment of the two English parse-trees. We can simplify this problem once we have a third parse tree (for the Chinese sentence) with node-to-node alignment relative to the two English trees. We will introduce such a loss function in the next section. We did not perform experiments involving this class of loss functions, but mention them for completeness in the hierarchy of loss functions.

2.3 Bilingual Parse-Tree Loss Functions

The third class of loss functions uses information from word strings, alignments and parse-trees in both languages, and can be described by

$$L((E, A), (E', A'); F) = L((T_E, A), (T_{E'}, A'); T_F).$$

We will now describe one such loss function using the example in Figures 1 and 2. Figure 3 shows a tree-to-tree mapping between the source (Chinese) parse-tree and parse-trees of its reference translation and two competing hypothesis (English) translations.

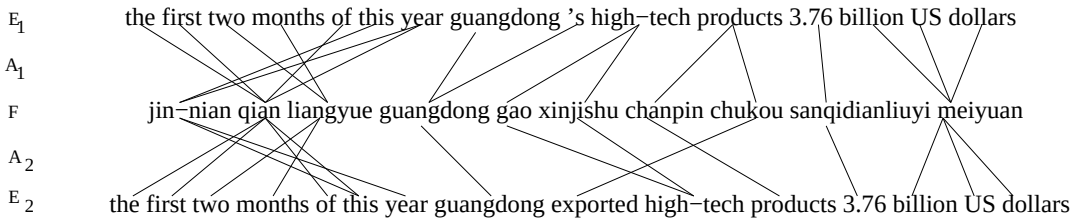

Figure 1: Two competing English translations for a Chinese sentence with their word-to-word alignments.

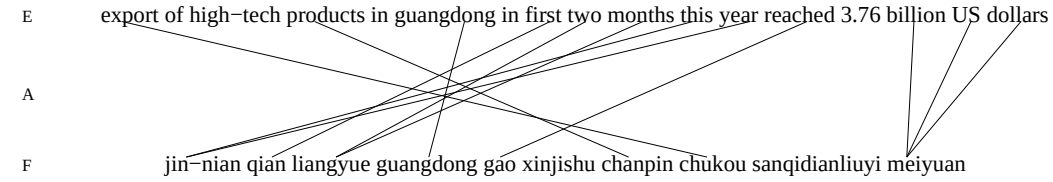

Figure 2: The reference translation for the Chinese sentence from Figure 1 with its word-to-word alignments. Words in the Chinese (English) sentence shown as unaligned are aligned to the NULL word in the English (Chinese) sentence.

We first assume that a node n in the source tree T_F can be mapped to a node m in T (and a node m' in T') using word alignment A (and A' respectively). We denote the subtree of T rooted at node m by t_m and the subtree of T' rooted at node m' by $t'_{m'}$. We will now describe a simple procedure that makes use of the word alignment A to construct node-to-node alignment between nodes in the source tree T_F and the target tree T.

2.3.1 Alignment of Parse-Trees

For each node n in the source tree T_F we consider the subtree t_n rooted at n. We first read off the source word sequence corresponding to the leaves of t_n. We next consider the subset of words in the target sentence that are aligned to any word in this source word sequence, and select the leftmost and rightmost words from this subset. We locate the leaf nodes corresponding to these two words in the target parse tree T, and obtain their closest common ancestor node $m \in T$. This procedure gives us a mapping from a node $n \in T_F$ to a node $m \in T$ and this mapping associates one subtree $t_n \in T_F$ to one subtree $t_m \in T$.

2.3.2 Loss Computation between Aligned Parse-Trees

Given the subtree alignment between T_F and T, and T_F and T', we first identify the subset of nodes in T_F for which we can identify a corresponding node in both T and T'.

$$\bar{N_F} = \{n \in T_F : m \neq \epsilon \cap m' \neq \epsilon\}.$$

The *Bilingual Parse-Tree (BiTree) Loss Function* can then be computed as

$$\text{BiTreeLoss}((T_E, A), (T_{E'}, A'); T_F) = \sum_{n \in \bar{N}_F} d(t_m, t'_{m'}), \quad (1)$$

where $d(t, t')$ is a distance measure between sub-trees t and t'. Specific Bi-tree loss functions are determined through particular choices of d. In our experiments, we used a 0/1 loss function between sub-trees t and t'.

$$d(t, t') = \begin{cases} 1 & t \neq t' \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

We note that other tree-to-tree distance measures can also be used to compute d, e.g. the distance function could compare if the subtrees t and t' have the same headword/non-terminal tag.

The Bitree loss function measures the distance between two trees in terms of distances between their corresponding subtrees. In this way, we replace the string-to-string (Levenshtein) alignments (for WER) or n-gram matches (for BLEU/PER) with subtree-to-subtree alignments.

The *Bitree Error Rate* (in %) is computed as a ratio of the Bi-tree Loss function to the number of nodes in the set $\bar{N}_F$.

The complete node-to-node alignment between the parse-tree of the source (Chinese) sentence and the parse trees of its reference translation and the two hypothesis translations (English) is given in Table 1. Each row in this table shows the alignment between a node in the Chinese parse-tree and nodes in the reference and the two hypothesis parse-trees. The computation of the Bitree Loss function and the Bitree Error Rate is presented in the last two rows of the table.

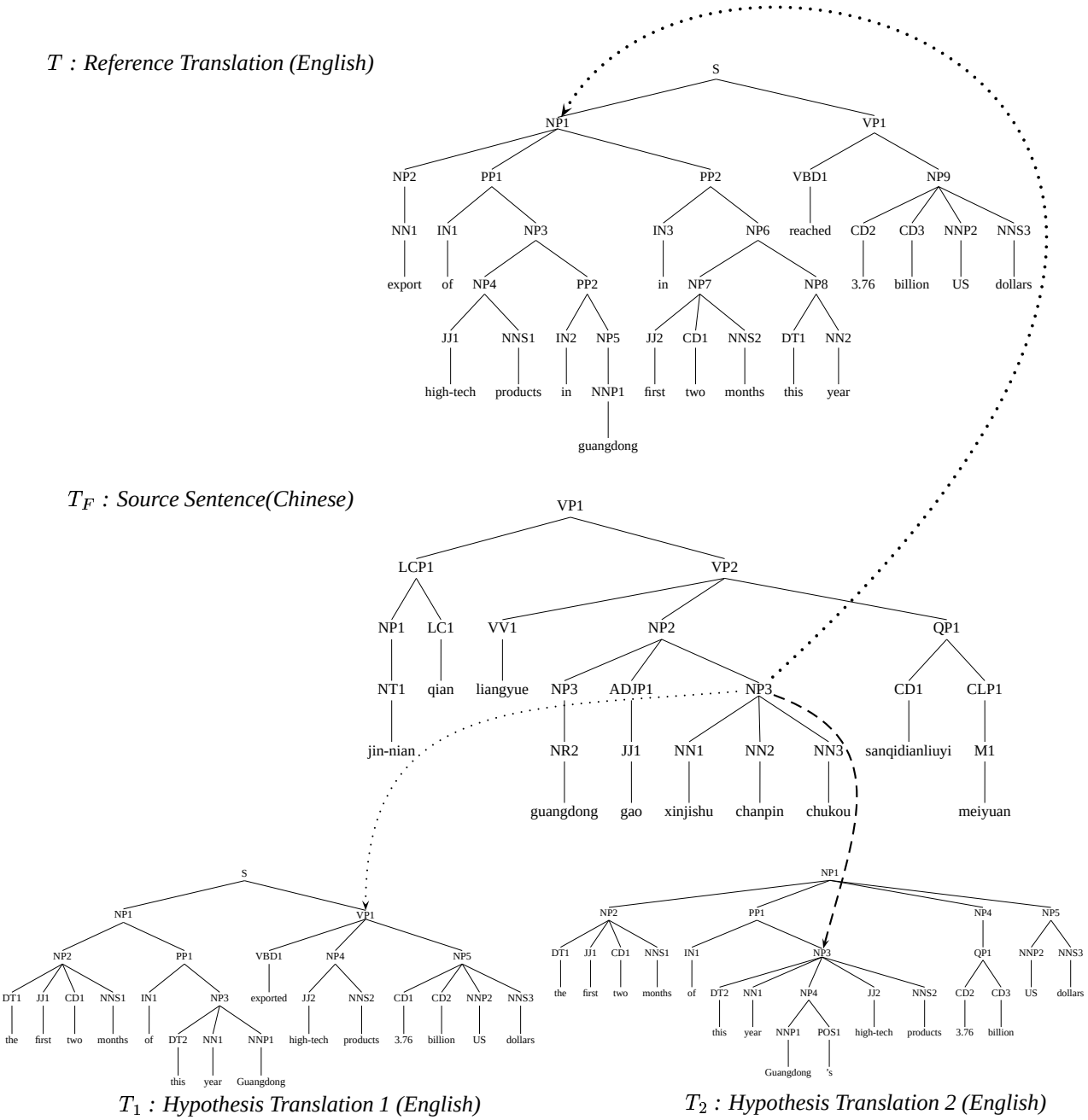

Figure 3: An example showing a parse-tree for a Chinese sentence and parse-trees for its reference translation and two competing hypothesis translations. We show a sample alignment for one of the nodes in the Chinese tree with its corresponding nodes in the three English trees. The complete node-to-node alignment between the parse-trees of the Chinese sentence and the three English sentences is given in Table 1.

Node $n \in T_F$	Node $m \in T$	Node $m_1 \in T_1$	$L(t_m, t_{m_1})$	Node $m_2 \in T_2$	$L(t_m, t'_{m_2})$
VP1	S	S	1	NP1	1
LCP	NP6	NP1	1	NP1	1
NP1	NP8	NP3	1	NP3	1
NT1	NP8	NP3	1	NP3	1
jin-nian	NP8	NP3	1	NP3	1
LC1	first	NP1	1	NP2	1
qian	first	NP1	1	NP2	1
VP2	S	S	1	NP1	1
VV	NP7	NP2	1	NP2	1
liangyue	NP7	NP2	1	NP2	1
NP2	S	S	1	NP1	1
NP3	Guangdong	Guangdong	0	NP4	1
NR2	Guangdong	Guangdong	0	NP4	1
guangdong	Guangdong	Guangdong	0	NP4	1
ADJP1	reached	high-tech	1	high-tech	1
JJ1	reached	high-tech	1	high-tech	1
gao	reached	high-tech	1	high-tech	1
NP3	NP1	VP1	1	NP3	1
NN2	products	products	0	products	0
chanpin	products	products	0	products	0
NN3	export	exported	1	products	1
chukou	export	exported	1	products	1
QP1	NP9	NP5	0	NP1	1
CLP1	NP9	NP5	0	NP1	1
M1	NP9	NP5	0	NP1	1
meiyuan	NP9	NP5	0	NP1	1
BiTree Loss		Loss(E, E_1)	17	Loss(E, E_2)	24
BiTree Error Rate (%)			17/26 = 65.4		24/26 = 92.3

Table 1: Bi-Tree Loss Computation for the parse-trees shown in Figure 3. Each row shows a mapping between a node in the parse-tree of the Chinese sentence and the nodes in parse-trees of its reference translation, hypothesis translation 1 and hypothesis translation 2.

2.4 Comparison of Loss Functions

In Table 2 we compare various translation loss functions for the example from Figure 1. The two hypothesis translations are very similar at the word level and therefore the BLEU score, PER and the WER are identical. However we observe that the sentences differ substantially in their syntactic structure (as seen from Parse-Trees in Figure 3), and to a lesser extent in their word-to-word alignments (Figure 1) to the source sentence. The first hypothesis translation is parsed as a sentence $S \rightarrow NP\ VP$ while the second translation is parsed as a noun phrase. The Bitree loss function which depends both on the parse-trees and the word-to-word alignments, is therefore very different for the two translations (Table 2). While string based metrics such as BLEU, WER and PER are insensitive to the syntactic structure of the translations, BiTree Loss is able to measure this aspect of translation quality, and assigns different scores to the two translations.

We provide this example to show how a loss function which makes use of syntactic structure from source and target parse trees, can capture properties of translations that string based loss functions are unable to measure.

Loss Functions	$L(E, E_1)$	$L(E, E_2)$
BLEU (%)	26.4	26.4
WER (%)	70.6	70.6
PER (%)	23.5	23.5
BiTree Error Rate (%)	65.4	92.3

Table 2: Comparison of the different loss functions for hypothesis and reference translations from Figures 1, 2.

3 Minimum Bayes-Risk Decoding

Statistical Machine Translation (Brown et al., 1990) can be formulated as a mapping of a word sequence F in a source language to word sequence E' in the target language that has a word-to-word alignment A' relative to F. Given the source sentence F, the MT decoder $\delta(F)$ produces a target word string E' with word-to-word alignment A'. Relative to a reference translation E with word alignment A, the decoder performance is measured as $L((E, A), \delta(F))$. Our goal is to find the decoder that has the best performance over all translations. This is measured through Bayes-Risk :

$$R(\delta(F)) = E_{P(E,A,F)}[L((E, A), \delta(F))].$$

The expectation is taken under the true distribution $P(E, A, F)$ that describes translations of human quality.

Given a loss function and a distribution, it is well known that the decision rule that minimizes the Bayes-Risk is given by (Bickel and Doksum, 1977; Goel and Byrne, 2000):

$$\delta(F) = \underset{E',A'}{\operatorname{argmin}} \sum_{E,A} L((E,A),(E',A');F)P(E,A|F). \tag{3}$$

We shall refer to the decoder given by this equation as the Minimum Bayes-Risk (MBR) decoder. The MBR decoder can be thought of as selecting a *consensus* translation: For each sentence F, Equation 3 selects the translation that is closest on an average to all the likely translations and alignments. The closeness is measured under the loss function of interest.

This optimal decoder has the difficulties of search (minimization) and computing the expectation under the true distribution. In practice, we will consider the space of translations to be an N-best list of translation alternatives generated under a baseline translation model. Of course, we do not have access to the true distribution over translations. We therefore use statistical translation models (Och, 2002) to approximate the distribution $P(E, A|F)$.

Decoder Implementation: The MBR decoder (Equation 3) on the N-best List is implemented as

$$\hat{i} = \underset{i\in\{1,2,\ldots,N\}}{\operatorname{argmin}} \sum_{j=1}^{N} L((E_j,A_j),(E_i,A_i))P(E_j,A_j|F)$$

and $\delta(F) = (E_{\hat{i}}, A_{\hat{i}})$. This is a rescoring procedure that searches for consensus under a given loss function. The posterior probability of each hypothesis in the N-best list is derived from the joint probability assigned by the baseline translation model.

$$P((E_j,A_j)|F) = \frac{P(E_j,A_j,F)}{\sum_{i=1}^{N} P(E_i,A_i,F)}. \tag{4}$$

The conventional Maximum A Posteriori (MAP) decoder can be derived as a special case of the MBR decoder by considering a loss function that assigns a equal cost (say 1) to all misclassifications. Under the 0/1 loss function,

$$L((E,A),(E',A')) = \begin{cases} 0 & \text{if } E = E' \,\&A = A' \\ 1 & \text{otherwise,} \end{cases} \tag{5}$$

the decoder of Equation 3 reduces to the MAP decoder

$$\delta_{\text{MAP}}(F) = \underset{(E',A')}{\operatorname{argmax}} P(E',A'|F). \tag{6}$$

This illustrates why we are interested in MBR decoders based on other loss functions: the MAP decoder is optimal with respect to a loss function that is very harsh. It does not distinguish between different types of translation errors and good translations receive the same penalty as poor translations.

4 Performance of MBR Decoders

We performed our experiments on the Large-Data Track of the NIST Chinese-to-English MT task (NIST, 2003). The goal of this task is the translation of news stories from Chinese to English. The test set has a total of 1791 sentences, consisting of 993 sentences from the NIST 2001 MT-eval set and 878 sentences from the NIST 2002 MT-eval set. Each Chinese sentence in this set has four reference translations.

4.1 Evaluation Metrics

The performance of the baseline and the MBR decoders under the different loss functions was measured with respect to the four reference translations provided for the test set. Four evaluation metrics were used. These were multi-reference Word Error Rate (mWER) (Och, 2002), multi-reference Position-independent word Error Rate (mPER) (Och, 2002) , BLEU and multi-reference BiTree Error Rate.

Among these evaluation metrics, the BLEU score directly takes into account multiple reference translations (Papineni et al., 2001). In case of the other metrics, we consider multiple references in the following way. For each sentence, we compute the error rate of the hypothesis translation with respect to the most similar reference translation under the corresponding loss function.

4.2 Decoder Performance

In our experiments, a baseline translation model (JHU, 2003), trained on a Chinese-English parallel corpus (NIST, 2003) ($170M$ English words and $157M$ Chinese words), was used to generate 1000-best translation hypotheses for each Chinese sentence in the test set. The 1000-best lists were then rescored using the different translation loss functions described in Section 2.

The English sentences in the N-best lists were parsed using the Collins parser (Collins, 1999), and the Chinese sentences were parsed using a Chinese parser provided to us by D. Bikel (Bikel and Chiang, 2000). The English parser was trained on the Penn Treebank and the Chinese parser on the Penn Chinese treebank.

Under each loss function, the MBR decoding was performed using Equation 3. We say we have a matched condition when the same loss function is used in both the error rate and the decoder design. The performance of the MBR decoders on the NIST 2001+2002 test set is reported in Table 3. For all performance metrics, we show the 70% confidence interval with respect to the MAP baseline computed using bootstrap resampling (Press et al., 2002; Och, 2003). We note that this significance level

does meet the customary criteria for minimum significance intervals of 68.3% (Press et al., 2002).

We observe in most cases that the MBR decoder under a loss function performs the best under the corresponding error metric i.e. matched conditions perform the best. The gains from MBR decoding under matched conditions are statistically significant in most cases. We note that the MAP decoder is not optimal in any of the cases. In particular, the translation performance under the BLEU metric can be improved by using MBR relative to MAP decoding. This shows the value of finding decoding procedure matched to the performance criterion of interest.

We also notice some affinity among the loss functions. The MBR decoding under the Bitree Loss function performs better under the WER relative to the MAP decoder, but perform poorly under the BLEU metric. The MBR decoder under WER and PER perform better than the MAP decoder under all error metrics. The MBR decoder under BLEU loss function obtains a similar (or worse) performance relative to MAP decoder on all metrics other than BLEU.

5 Discussion

We have described the formulation of Minimum Bayes-Risk decoders for machine translation. This is a general framework that allows us to build special purpose decoders from general purpose models. The procedure aims at direct minimization of the expected risk of translation errors under a given loss function. In this paper we have focused on two situations where this framework could be applied.

Given an MT evaluation metric of interest such as BLEU, PER or WER, we can use this metric as a loss function within the MBR framework to design decoders optimized for the evaluation criterion. In particular, the MBR decoding under the BLEU loss function can yield further improvements on top of MAP decoding.

Suppose we are interested in improving syntactic structure of automatic translations and would like to use an existing statistical MT system that is trained without any linguistic features. We have shown in such a situation how MBR decoding can be applied to the MT system. This can be done by the design of translation loss functions from varied linguistic analyzes. We have shown the construction of a Bitree loss function to compare parse-trees of any two translations using alignments with respect to a parse-tree for the source sentence. The loss function therefore avoids the problem of unconstrained tree-to-tree alignment. Using an example, we have shown that this loss function can measure qualities of translation that string (and ngram) based metrics cannot capture. The MBR decoder under this loss function gives improvements under an evaluation metric based on the loss function.

We present results under the Bitree loss function as an example of incorporating linguistic information into a loss function; we have not yet measured its correlation with human assessments of translation quality. This loss function allows us to integrate syntactic structure into the statistical MT framework without building detailed models of syntactic features and retraining models from scratch. However, we emphasize that the MBR techniques do not preclude the construction of complex models of syntactic structure. Translation models that have been trained with linguistic features could still benefit by the application of MBR decoding procedures.

That machine translation evaluation continues to be an active area of research is evident from recent workshops (AMTA, 2003). We expect new automatic MT evaluation metrics to emerge frequently in the future. Given any translation metric, the MBR decoding framework will allow us to optimize existing MT systems for the new criterion. This is intended to compensate for any mismatch between decoding strategy of MT systems and their evaluation criteria. While we have focused on developing MBR procedures for loss functions that measure various aspects of translation quality, this framework can also be used with loss functions which measure application-specific error criteria.

We now describe related training and search procedures for NLP that explicitly take into consideration task-specific performance metrics. Och (2003) developed a training procedure that incorporates various MT evaluation criteria in the training procedure of log-linear MT models. Foster et al. (2002) developed a text-prediction system for translators that maximizes expected benefit to the translator under a statistical user model. In parsing, Goodman (1996) developed parsing algorithms that are appropriate for specific parsing metrics. There has also been recent work that combines 1-best hypotheses from multiple translation systems (Bangalore et al., 2002); this approach uses string-edit distance to align the hypotheses and rescores the resulting lattice with a language model.

In future work we plan to extend the search space of MBR decoders to translation lattices produced by the baseline system. Translation lattices (Ueffing et al., 2002; Kumar and Byrne, 2003) are a compact representation of a large set of most likely translations generated by an MT system. While an N-best list contains only a limited reordering of hypotheses, a translation lattice will contain hypotheses with a vastly greater number of re-orderings. We are developing efficient lattice search procedures for MBR decoders. By extending the search space of the decoder to a much larger space than the N-best list, we expect further performance improvements.

MBR is a promising modeling framework for statistical machine translation. It is a simple model rescoring framework that improves well-trained statistical models

	Performance Metrics			
Decoder	BLEU (%)	mWER(%)	mPER (%)	mBiTree Error Rate(%)
70% Confidence Intervals	+/-0.3	+/-0.9	+/-0.6	+/-1.0
MAP(baseline)	31.2	64.9	41.3	69.0
MBR				
BLEU	**31.5**	65.1	41.1	68.9
WER	31.3	**64.3**	40.8	68.5
PER	31.3	64.6	**40.4**	68.6
BiTree Loss	30.7	64.1	41.1	**68.0**

Table 3: Translation performance of the MBR decoder under various loss functions on the NIST 2001+2002 Test set. For each metric, the performance under a matched condition is shown in bold. Note that better results correspond to higher BLEU scores and to lower error rates.

by tuning them for particular criteria. These criteria could come from evaluation metrics or from other desiderata (such as syntactic well-formedness) that we wish to see in automatic translations.

Acknowledgments

This work was performed as part of the 2003 Johns Hopkins Summer Workshop research group on *Syntax for Statistical Machine Translation*. We would like to thank all the group members for providing various resources and tools and contributing to useful discussions during the course of the workshop.

References

AMTA. 2003. Workshop on Machine Translation Evaluation, MT Summit IX. www.issco.unige.ch/projects/isle/MTE-at-MTS9.html.

S. Bangalore, V. Murdock, and G. Riccardi. 2002. Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of COLING*, Taipei, Taiwan.

P. J. Bickel and K. A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected topics*. Holden-Day Inc., Oakland, CA, USA.

D. Bikel and D. Chiang. 2000. Two statistical parsing models applied to the chinese treebank. In *Proceedings of the Second Chinese Language Processing Workshop*, pages 1–6, Hong Kong.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

M. Collins and N. Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the weighted perceptron. In *Proceedings of EMNLP*, Philadelphia, PA, USA.

M. J. Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of HLT 2002*, San Diego, CA. USA.

G. Foster, P. Langlais, and G. Lapalme. 2002. User-friendly text prediction for translators. In *Proc. of EMNLP*, Philadelphia, PA, USA.

V. Goel and W. Byrne. 2000. Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 14(2):115–135.

J. Goodman. 1996. Parsing algorithms and metrics. In *Proc. of ACL-1996*, pages 177–183, Santa Cruz, CA, USA.

JHU. 2003. Syntax for statistical machine translation, Final report, JHU summer workshop. http://www.clsp.jhu.edu/ws2003/groups/translate/.

S. Kumar and W. Byrne. 2002. Minimum Bayes-Risk alignment of bilingual texts. In *Proc. of EMNLP*, Philadelphia, PA, USA.

S. Kumar and W. Byrne. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of HLT-NAACL*, Edmonton, Canada.

I. D. Melamed, R. Green, and J. P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of the HLT-NAACL*, Edmonton, Canada.

NIST. 2003. The NIST Machine Translation Evaluations. http://www.nist.gov/speech/tests/mt/.

F. Och. 2002. *Statistical Machine Translation: From Single Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Germany.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, Sapporo, Japan.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

M. Tang, X. Luo, and S. Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of ACL 2002*, Philadelphia, PA, USA.

N. Ueffing, F. Och, and H. Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. of EMNLP*, pages 156–163, Philadelphia, PA, USA.